融合统计学习和语义过滤的 ADR 信号抽取模型构建研究*

■ 魏巍[1] 郑杜[2]
[1] 中南财经政法大学大数据研究院 武汉 430074 [2]武汉大学信息管理学院 武汉 430072

摘要：［目的/意义］社交媒体的出现为医疗健康数据的收集提供了新的途径，应用自然语言处理技术从社交媒体中抽取患者报告的 ADR(Adverse Drug Reaction，药物不良反应)信号对于改善药物不良反应监测的临床和科学知识具有很大的潜力。然而，从社会媒体中提取患者报告的 ADR 信号仍然面临重大挑战。为此，开发一个利用高级自然语言处理技术从健康主题社交媒体中抽取 ADR 信号的研究模型。［方法/过程］该模型首先采用基于多词典源匹配的方法，从嘈杂的社交媒体中识别医学实体；然后采用最短依存路径核函数为基础的统计学习方法提取药物不良事件；并利用药品安全数据库的语义知识过滤药物的治疗和适用症信息以及否定的药物不良事件；最后，对报告源进行分类剔除传闻等噪音信息。［结果/结论］通过收集糖尿病论坛上的数据对模型的有效性进行验证，结果显示该模型的每一部分都有助于其整体性能的提升。
关键词：医学实体识别 药物不良事件抽取 健康社交媒体 统计学习 语义过滤
分类号：G251
DOI:10.13266/j.issn.0252-3116.2018.05.013

1 引言

近年来，随着互联网和以 Web 2.0 为基础的社交媒体的快速发展，人们获取健康信息的方式正逐渐发生变化。由过去与医护人员面对面被动地交流医疗信息，到如今通过健康主题社交媒体主动搜索获取并分享健康信息，人们希望能够参与到自己健康的日常管理中。越来越多的患者愿意在互联网上尤其是网络健康社区中，分享他们的诊断、治疗、药物和副作用信息，以及自己在与疾病抗争中的情感经历[1]。这使得此类社交媒体成为独特和强大的获取健康、药物和治疗信息的重要来源。患者在社交媒体的自述，经常会包含一些临床医生可能错过或忽视的医疗问题和不良反应，网络健康社区中积累的评论信息是来自用药者的第一手资料，其中蕴含了丰富的潜在 ADR 信息。社交媒体被认为是一种收集药物副作用和治疗效果的新渠道，它能增强获取药品安全和治疗管理的主观要素，为临床实践提供重要见解。

鉴于社交媒体上患者报告内容的临床和科学价值，研究人员已经开始探索从社交媒体中识别和提取它们的方法[2]。社交媒体包含大量患者口语化的表达，从这种嘈杂的环境中提取高质量的患者报告内容是具有挑战性的。药物不良事件是由药物引起的医学事件，通常在患者的网络叙述中，治疗信息和医学事件经常混杂在一起出现，在他们的讨论中可能包含药物的治疗信息、适应症信息，以及否定的药物不良事件。药品适应症是药品使用的基本常识，是药物用于治疗疾病的合理的医学说明。否定的药物不良事件是对药物与不良事件之间因果关系的否定。社交媒体上的药物不良事件可能来自患者的真实经历，也可能是科研人员的研究、新闻、传闻或复制等信息，这就导致报告源中含有大量的噪音和重复数据[3]。表 1 通过糖尿病网络社区的帖子对以上讨论的现象进行了解释。

从表 1 列举的帖文中可以发现，网络社区用户在讨论中使用他们偏好的医学保健语言，这些语言不同于医学专业术语。例如，编号为 63828 的帖子中，“*stroke*”(中风)是用户偏好的表达，而在 FAERS(美国食品药品监督管理局的不良事件报告系统)中的术语表述为“*cerebrovascular*”(脑血管疾病)；34188 号帖子

* 本文系国家自然科学基金项目“基于文本和 web 语义分析的智能咨询服务研究”(项目编号：71673209)研究成果之一。
作者简介：魏巍(ORCID:0000-0003-3580-8360)，讲师，博士，E-mail:503175355@qq.com；郑杜，博士研究生。
收稿日期:2017-09-07 修回日期:2017-12-05 本文起止页码:115-124 本文责任编辑:王善军

图书情报工作
第 62 卷 第 5 期 2018 年 3 月

表 1 用户在社交媒体上生成内容的实例

帖子 ID	贴文内容	是否包含 ADE	报告源
9043	I had horrible chest pain [Event] under Actos[Treatment]	ADE	患者报告
12200	From what you have said, it seems that Lantus [Treatment]has hadsome negative side effects related to depression [Event] and mood swings [Event]	ADE	传闻
25139	I never experienced fatigue [Event] whenusing Zocor [Treatment]	No	患者报告
34188	When taking Zocor [Treatment], I hadheadaches [Event] and bruising [Event]	ADE	患者报告
63828	Another study of people with multiple riskfactors for stroke [Event] found thatLipitor [Treatment] reduced the risk of stroke [Event] by 26% compared to those taking a placebo, the company said	药品适用症	糖尿病研究

中,“*bruising*”(擦伤)在 FAERS 中的表达为“*contusion*”(擦伤);此外,患者在讨论中可能包含不同类型的药物和不良事件关系:例如在 63828 号帖子中,笔者提到了“*stroke*”和“*Lipitor*”(立普妥),“*Lipitor*”是降低中风风险的降脂剂,“*stroke*”和“*Lipitor*”在这篇帖子中呈现的是药物适应症的关系,而不是药物 – 不良反应关系;而在 9043 号帖子中,患者报告了在服用“*Actos*”(艾可拓,一种降血糖药)时有胸痛现象,呈现为药物不良事件。论坛中的信息还可能来自不同的报告源,如 63828 号帖子是关于糖尿病的研究,9043、25139、34188 号帖子是患者亲身经历的用药评论,而 12200 号帖子为从别人那里听到的传闻。

2 相关研究

医学实体识别旨在确定医疗实体对象,如治疗和药品等。归功于医疗健康领域丰富的医学词典及知识库,以往的许多研究都采用基于词典的实体识别方法。UMLS(美国国家医学图书馆开发的一体化医学语言系统)在研究中常被采用[4];自发报告系统也经常被用来作为从文本中提取治疗和不良事件的数据源;FAERS 的医学术语常被用于映射健康社交媒体的药物和不良事件实体[5];MedEffect(加拿大的药物不良事件报告系统)也被用来从社交媒体提取不良事件[6]。然而,健康社交媒体上用户生成的保健用语是不同于医学专业术语的,网络用户由于个人知识和偏好的差异,对药物评论的表达亦可能是五花八门的。

从预处理后的数据中识别医学实体的方法包括基于规则的方法、基于词典的方法和基于统计学习的方法。在实际应用中通常会根据具体任务的要求,选择某一种或几种方法以期获得更好的识别效果[7]。S. Abeed 等[8]的文献调查显示,药物不良反应词典和知识库一直是利用社交媒体进行 ADR 信号抽取广泛使用的数据资源。这些生物医学数据资源中包含了 ADR 列表,收集了从药品标签到临床试验、看护者,甚至社交媒体上用户的帖子等数据内容。

生物医学关系抽取技术已经被用于从自由文本中鉴别诸如基因 – 疾病关系以及蛋白质的相互作用关系等。药物不良事件的抽取采用关系抽取技术来确定药物和事件之间是否存在关系以及关系的类型(例如药物 – 适应症关系或者药物 – 不良反应关系)。药物不良反应关系的抽取方法可以分为三类:基于共现分析(co-occurrence analysis)的方法、基于规则(rule-based)的方法和基于统计学习(statistical learning based)的方法。对于每种方法,不同之处在于如何更好地利用文本中的词汇、语法和语义信息。基于规则的句法和语义信息抽取方法表现出较好的性能;基于统计学习的关系抽取方法可以从标注的语料库中自动地学习关系模式,更适用于大规模语料的需求。有监督的统计学习在实体关系抽取中占据主导位置。其中,基于核函数的实体关系抽取就是一种有代表性的方法。基于核函数的关系抽取方法在确定各种生物医学关系如蛋白质相互作用和基因 – 疾病的关系时,已显示出它的优势。P. Thomas 等采用复合核函数,集成了学习图形核和最短依存路径核函数从医学文献中提取药物 – 药物相互作用关系[9]。它们利用基于句法和语义的信息,能够更简洁、准确地捕获实体之间的关系,从而比基于特征的关系抽取方法获得了更好的效果。这种方法利用核函数可以将多方面的语法、语义等信息综合,最终的实体关系距离由多个不同信息来源的核函数复合而成,从而可以提高准确率[10]。

社交媒体上大量在线用户的口语化生成内容,将会导致大量的稀疏性词汇特征集,从而使得基于特征的关系抽取方法的性能大幅降低。然而,健康社交媒体的用户讨论仍然遵循一定的语法和语义模式。基于核函数的统计学习方法借助数据之间的句法和语义表示,可用于从嘈杂的社交媒体文本中提取药物不良反应关系。

以往的大多数研究都使用准确率、召回率和 F 值指标来评估其性能的优劣。为了证明来自社交媒体报告的药物不良反应的价值，研究人员对提取的结果进行了多项分析。A. Benton 等[11]将从社交媒体提取的不良事件与已记录的药物不良事件进行比较，发现与记录的药物不良事件相比，从社交媒体中提取的药物不良事件可以达到 35.1% 的准确率、77% 的召回率和 52.8% 的 F 值；B. Chee 等[6]发现患者的药物评论可以用来识别市场上的风险药物，并且识别出来的大多数风险药物都出现在美国食品和药品监督管理局的药品安全观察名单上；C. Yang 等[12]认为健康社交媒体是 ADR 信号检测中具有广阔应用前景的重要数据源。

通过对已有研究的回顾，笔者发现基于医学词典和本体的医学实体抽取能够达到令人满意的效果[13]。利用高级的统计学习进行关系抽取的方法，应用在健康社交媒体挖掘药物不良反应的研究较少。基于共现分析的药物不良反应提取方法存在明显的局限性：这种方法不能很好地捕获语法或语义信息，其结果导致当句子中存在否定关系时，可能提取的不良反应关系是错误的；提取的药物不良事件可能与药物的适应症相互混淆；当多个药物不良反应实体同时出现在同一个句子中时，这种方法无法准确地捕获药物与不良反应之间的关系。此外，健康社交媒体中有许多来自第三方账号的新闻、研究、故事等的重复报告，这些内容会产生冗余和噪音，从而降低社交媒体识别 ADR 信号的准确率，之前的研究很少关注这个问题。健康社交媒体作为一个日受瞩目的开放平台，用户通过 Web 社区可以自由地说出自己的问题和诉求，其价值还远远没有得到充分的挖掘。

本文提出的研究问题可以描述为：开发一个集成和可扩展的研究模型用于从健康主题社交媒体中挖掘患者报告的 ADR 信号；从嘈杂的健康主题社交媒体的用户讨论中识别出真正的患者报告内容；与基准方法相比，统计学习方法需要增强健康相关的语义过滤来改善药物不良事件提取的结果。

3　融合统计学习和语义过滤的 ADR 信号抽取模型

鉴于利用社交媒体进行药物不良反应监测的研究价值以及当前从用户生成内容中提取药物不良事件的障碍，笔者提出了一个从健康主题社交媒体中抽取患者报告的 ADR 信号的研究模型，在这个模型中，设计了一个基于多词典源的医学实体抽取方法，它集成了多个医学词典和网民保健用语来解释用户生成的口语化的医学健康语言。此外，该模型使用基于最短依存路径核函数的统计学习方法和基于医学知识库信息的语义过滤方法进行药物不良事件关系的提取。这种方法利用现有的医学知识和统计学习技术，可以显著地增强提取的不良事件的准确率。为了将真实的患者报告从第三方转载中识别出来，还对报告来源进行分类，以识别患者真正报告的药物不良事件。该模型包括数据的预处理、基于多词典源的医学实体识别、基于最短依存路径核函数和语义过滤的药物－不良事件关系抽取、以及对报告源进行分类 4 个组成部分，如图 1 所示：

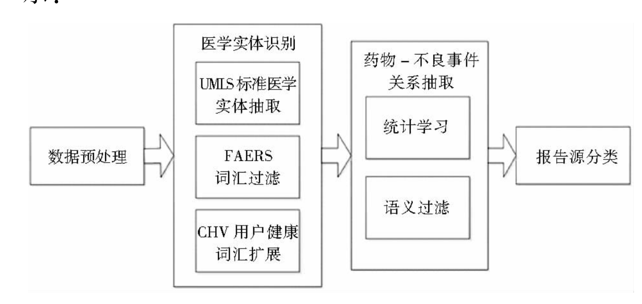


图 1　面向健康主题社交媒体的 ADR 信号抽取模型

3.1　数据收集及预处理

数据预处理通常是对数据进行清洗和标准化，为后续分析准备原始数据。数据预处理阶段包括两个步骤：文本清洗和断句。基于正则表达式，去除 URL、重复的标点符号和文本中的个人身份信息，如电子邮件地址、个人账号、电话号码，剔除掉无关信息的同时保留有用信息以确保迭代过程的速度和结果的质量。笔者提出的方法集中在句子层面信息的提取和处理，因此，使用自然语言处理工具 OpenNLP 对每篇贴文进行断句。OpenNLP 提供最新的基于机器学习的句子边界检测算法，利用它将爬取的帖文分割成独立的句子。

3.2　基于多词典源的医学实体识别

从嘈杂的用户生成内容中提取医疗实体是一件具有挑战的任务。R. Leaman 等[4]基于词典方法的研究被证明是表现最佳的医疗实体识别系统。基于词典的方法依赖于现有的词典，通常是基于字符串匹配或相似度计算从自由文本中识别药物和不良事件实体。这种识别方法的性能取决于底层参照词典的全面性及相似度算法的优劣。笔者将利用 UMLS、FEARS 和 CHV

图书情报工作
第 62 卷 第 5 期　2018 年 3 月

多词典源从社交媒体文本中抽取药名和药物不良事件实体。

3.2.1　基于 UMLS 的标准医学实体抽取　MetaMap 是一个链接美国国家医学图书馆的 java API,用于从健康社交媒体识别 UMLS 医学概念。目前,UMLS 有 135 种语义类型,这些类型被进一步抽象成 15 个语义组,如"Chemicals and Drugs""Disorders""Genes & Molecular Sequences"等。通过配置 MetaMap,识别属于"Chemicals and Drugs"语义组的药名实体和属于"Disorders"语义组的药物不良事件实体。首先通过 MetaMap 鉴别患者评论中与标准医学词典 UMLS 匹配的医学实体。

3.2.2　基于 FAERS 的医学词汇过滤　MetaMap 映射"Chemicals and Drugs"语义组以及"Disorders"语义组的结果中可能包含一些错误的正例信息。例如,在论坛讨论中的食品和配方成分通常被认定属于"Chemicals and Drugs"语义组;常见的动词,如"find"和"have"可能被提取为"Disorders"语义组。为了避免这些问题,利用 FDA 的 FAERS 对从 MetaMap 提取的药名和不良事件名进行筛选,剔除那些未在 FEARS 中出现的医学实体名,待作进一步分析。

3.2.3　基于 CHV 的网民保健用语扩充　网络健康社区中患者讨论的医学问题不同于医学文献,论坛中的用户生成内容通常包含用户偏好的医学词汇和描述性文本。为了更全面准确地了解社交媒体中的患者讨论内容,笔者集成网民保健用语(Consumer Health Vocabulary,CHV)作为词典源,扩展更为丰富的在线用户表达。CHV 中包含 47 505 个 UMLS 标准医学术语和对应的 127 081 个用户偏好词汇。对前面保留下来的每一个医学实体,查询 CHV 得到其对应的用户偏好词汇,这些偏好词汇之前是无法被 MetaMap 识别的,然后利用这些用户偏好术语来检索患者评论数据集,以扩展医学实体的抽取。将网络健康社区中提及的用户偏好词汇集成,进一步扩充了抽取的医学实体集。

3.3　基于最短依存路径核函数的药物不良事件关系抽取

网络健康社区中患者讨论的药物不良事件不同于生物医学文献或临床笔记,这些讨论通常包含更多非正式的和口语化的表达,这需要医学知识和复杂的语言技术进行解析。通过前文中对生物医学关系抽取研究的回顾,结合关系检测的统计学习方法和基于医学及语言知识库的语义信息过滤方法来识别药物不良事件。

基于最短依存路径核函数的实体关系抽取方法,首先以句子为单位,列出句中的所有实体对,为每一实体对建立一个依存树。依存树描述了句子中实体间的语法关系,如主语和它所支配动词的依存关系,形容词和它所修饰名词的依存关系。然后基于依存树设计核函数,通过核函数计算实体关系的距离,最后用支持向量机将数据分类。

基于最短依存路径核函数的药物 - 不良事件关系抽取可以确定一个句子中的药物和医学事件是否存在关系。本文的研究模型中开发了一个基于最短依存路径核函数的统计学习方法。最短依存路径核函数在识别各种关系(如基因相互作用和药物相互作用等)方面已显示出它的优势[9,14]。笔者将利用最短依存路径核函数和支持向量机(SVM)从药物不良反应相关的帖子中获得学习模式提取药物不良事件。基于最短依存路径核函数的药物不良事件关系抽取方法主要包括特征生成,核函数和分类三个部分,如图 2 中上半部分所示:

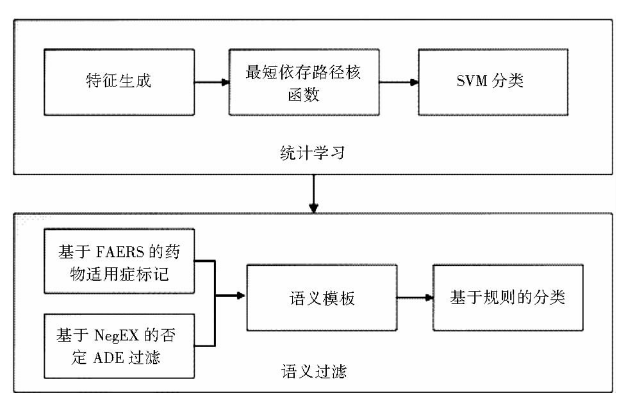


图 2　药物 - 不良事件关系抽取过程

3.3.1　特征生成　网络健康社区中患者谈论的药物不良事件通常包含大量的口语化表达,由于数据的稀疏性,以词汇和距离为特征的统计学习效果不令人满意。然而,患者对药物不良事件的叙述仍然遵循某些句法和语义规则,因此建议从句子依存解析树中提取句法和语义特征来表示实例。依存解析基于句法关系生成词到词的链接,它们表示句子中词汇间的语法和语义信息。在依存解析树中,句法依存性会显示在树的层次结构中,语义依存性会在链接的方向中显示。采用 Stanford Parser 进行依存句法解析从依存者到支配者的语法关系。Stanford Parser 运用上下文无关文

魏巍，郑杜.融合统计学习和语义过滤的 ADR 信号抽取模型构建研究[J].图书情报工作,2018,62(5):115 - 124.

法和词汇化依存句法分析,生成依存树中各成分之间的依存关系。图 3 是一个句子的依存关系树。在这个句子中,“nausea”是不良事件实体,“Byetta”是治疗糖尿病的一种药物。图中示出了词间的语法关系。例如,“nausea”是“gotten”的直接对象,因此它们具有语法关系“dobj”。在这种情况下,“gotten”是支配者,“nausea”是依存者。

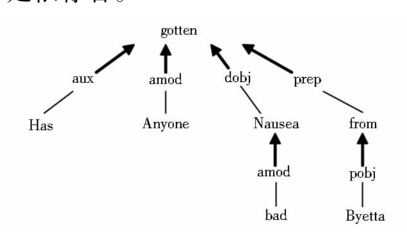


图 3　一个句子的依存树表示实例

依存树的大部分是与句子中的药物和医疗状况关系无关的。之前已有研究表明,依存树在确立两个实体之间关系的贡献几乎完全集中在其上的两个实体之间的最短路径。为了利用药物实体和医疗事件实体在依存关系中的最短路径(最短依存路径),提出算法从依存树中抽取两个实体的最短路径。算法针对每个关系实例在依存树中搜索从医疗事件到药物治疗的最短路径,不仅捕获单词还包括路径上依存关系的方向。最短依存路径抽取的过程如伪码 1 所示:

伪码 1:

输入:一个关系实例 i,一对相关的药物和不良事件 $R(drug,\ event) = True$,依存图 T

输出:路径,从事件到药物的最短依存路径

步骤:最短依存路径抽取($i,\ drug,\ event,\ T$)

```
if drug event. dependents( ) then
    Path← {event. ←,drug}
else
    Path ←{event} ,End ←{drug} ,Head ←{event} ,Tail ←{drug}
    whileHead ≠Tail. governor do
    if drug Head . dependents( ) then
        Head ← Head. governor
        Path← Path + {→. Head,← ,drug}
    else
        Head← Head. governor
        Path← Path + {→. Head}
    if event Tail. governor. dependents( ) then
        Tai l← Tail. governor
        End←{event,→,Tail,←} + End
    else
        Tail ← Tail. governor
        End ← {Tail,←} + End
        Path← Path + {←} + End
        return path
```

3.3.2　语法和语义类映射　为了增加抽取方法的鲁棒性,将路径上的单词进行词性标注(POS tags)来扩展最短依存路径。在对句子进行浅层句法分析后,采用 Stanford CoreNLP 软件包抽取词性信息并进行标注,利用 Stanford Penn Tree Bank guidelines 生成词性标注;语义类型(事件和药物治疗)将被标记于最短路径的两端。表 2 列出了数据集中涉及的词性标注。

表 2　词性标注注释

Part-of-Speech (POS) tags	解释
CC	Conjunction (连词)
CD	Cardinal number(基数)
DT,PDT	Determiner(限定词)
IN	Preposition(介词)
JJ,JJR,JJS	Adjective(形容词)
NN,NNS,NNP,NNPS	Noun(名词)
PRP,PRPS	Pronoun(代词)
RB,RBR,RBS	Adverb(副词)
RP	Particle(小品词)
UH	Interjection(感叹词)
VB,VBD,VBG,VBN,VBZ,VBP	Verb(动词)
WDT,WP,WPS,WRB	Wh-words(特殊疑问词)
EX,FW,LS,MD,SYM	其它

句法分析树中的节点定义匹配模式与核函数。其中,匹配模式反映两个节点的词性和特征是否匹配;通过核函数进行关系实例的匹配与相似性计算,从而得出两个实体之间的关系。关系实例的特征表示可以定义为路径上所有元素的笛卡尔积。图 3 中示例句子的特征表示为如下等式。原始句子可以以一个序列表示为 $X = \{x_1, x_2, x_3, x_4, x_5\}$,其中,$x_1 = \{Nausea, NN, Noun, Event\}$,$x_2 = \{\rightarrow\}$,$x_3 = \{gotten,\ VBD, Verb\}$,$x_4 = \{\leftarrow\}$,$x_5 = \{Byetta, NN, Noun, Treatment\}$。

$$\begin{bmatrix} Nausea \\ NN \\ Noun \\ Event \end{bmatrix} \times [\rightarrow] \times \begin{bmatrix} gottten \\ VED \\ Verb \end{bmatrix} \times [\leftarrow] \times \begin{bmatrix} Byetta \\ NN \\ Noun \\ Treatment \end{bmatrix}$$

3.3.3　最短依存路径核函数　基于核函数的方法可以利用多种不同的数据组织形式表示实体关系。在计算关系之间的距离时,不再使用特征向量的内积而是

使用核函数。核函数在高维度特征空间中隐式地计算对象的特征向量的点积,也就是说,在许多情况下不用枚举出所有的特征也可以计算出它们所在位置共同特征的数量乘积。统计学习方法依赖于核函数找到一个超平面将正实例与负实例分离。对于最短依存路径核函数,如果 $x = x_1x_2x_3x_4...x_m$, $y = y_1y_2y_3y_4...y_n$ 是两个关系实例,其中 x_i 表示对应于位置 i 的特征集合,核函数的定义如公式(1):

$$K(x,y) = \begin{cases} 0 & m \neq n \\ \prod_{i=1}^{n} C(x_i, y_i) & m = n \end{cases} \quad \text{公式(1)}$$

$C(x_i, y_i) = |x_i \cap y_i|$ 是 x_i 和 y_i 之间的共同特征的数量。两个关系实例是否具有相同的关系类型,可以通过核函数计算得出。最短依存路径核函数的伪码描述如下:

伪码 2:

输入:关系实例 $x = x_1x_2 \cdots x_n$ 和 $y = y_1y_2 \cdots y_n$

输出:$K(x, y)$,x 与 y 的相似性得分

步骤:最短依存路径核函数(x, y)

1. If $m \neq n$ then
2. $K(x, y) \leftarrow 0$
3. else
4. whilei$\leq m$ do
5. $K(x, y) \leftarrow K(x, y)\ |x_i \cap y_i|$
6. returnK(x, y)

举例说明通过核函数进行关系实例的匹配与相似性计算,如一个关系实例 x = {*When this happens, the basal action of your Lantus could cause hypoglycemia.*},特征表示为 x = [{*Hypoglycemia*, *NN*, *Noun*, *Event*}, {→}, {*cause*, *VB*, *Verb*}, {←}, {*action*, *NN*, *Noun*}, {←}, {*Lantus*, *NN*, *Noun*, *Treatment*}];另一个关系实例 y = {*But, now I've read a few posts in this thread that indicate depression as a possible side effect from Lantus.*}可以表示为 y = [{*depression*, *NN*, *Noun*, *Event*}, {→}, {*indicate*, *VBP*, *Verb*}, {←}, {*effect*, *NN*, *Noun*}, {←}, {*Lantus*, *NNP*, *Noun*, *Treatment*}]。核函数 $K(x, y)$的计算为位置 i 中 x_i 和 y_i 共同特征的数量乘积。本例中,$K(x, y) = 3*1*1*2*1*3 = 18$。根据这个结果,可以得出两个关系实例 x 和 y 具有非常高的相似性得分。如果关系实例 x 具有某种药物-事件关系,则关系实例 y 很可能也包含该药物-事件关系。

3.3.4 分类 实体关系检测中的分类旨在将具有某种关系特征的关系实例与不具备该种关系的实例区分开。采用直推式支持向量机(TSVM)用于关系检测的分类。SVM-light 是支持直推式支持向量机的开源软件包,在之前的研究中已被采用并取得较好的效果,更重要的是它还具有用户自定义核函数的功能[15]。通过 SVM-light 自定义最短依存路径核函数,根据最短依存路径核函数训练 TSVM 分类器,然后应用这个分类器来识别药物-事件关系的实例。详细过程如伪码 3 所示:

伪码 3:

输入:所有关系实例 I,每个实例至少包含一对药物和事件

输出:是否一对药物和不良事件是相关的

$R\,(drug,\ event) = True\ or\ False$

步骤:统计学习算法$(drug,\ event)$

1. 对于每一对药物和事件,$R\,(drug,\ event)$ do
2. 生成包含 $R\,(drug,\ event)$实例 i 的依存图 T
3. $Path$ ←最短依存路径抽取 $R\,(i,\ drug,\ event,\ T)$
4. $Feature$←语法与语义类匹配$(Path)$
5. 将关系实例分为训练集和测试集
6. 在训练集上利用最短依存路径核函数训练一个 SVM 分类器 C
7. 在测试集上使用分类器 C 将关系实例分为两类:

$R\,(drug,\ event) = True$

$R\,(drug,\ event) = False$

3.4 语义过滤

最短依存路径核函数可以检测到相关的药物和不良反应关系,然而,该方法还不能精确地捕获句子中的否定关系,也不能将药物适应症与药物不良反应区分开。已有的研究都忽视了过滤掉药品适应症和否定的药物不良反应进行分析的重要性,导致抽取的药物不良事件准确率偏低。为了解决这个题,笔者采用一个语义过滤方法,基于药品安全数据库的语义知识,过滤掉药品适应症信息,并利用否定检测工具中的规则过滤掉否定的药物不良反应信息(见图 2 中下半部分所示)。

3.4.1 基于 FAERS 的药物适用症标记 患者在社区中分享用药经历或评论某种药物时,不可避免的会提到用药原因或药物的适应症。比如药品 *Metoprolol* 中的一条评论:"*I use this primarily for my hypertension*。"这句话表达的意思中,"*hypertension*"是用药的原因,而不是 *Metoprolol* 的不良反应。由于药物适应症是规范化的,并且在药品安全数据库(例如 FAERS)中有详细

的记录，因此，从 FAERS 中获取药物适应症知识并形成模板，利用 MetaMap 从 FAERS 的适应症描述中识别出相关的生物医学实体，将混杂在药物不良事件中的药品适应症信息过滤掉。

3.4.2　基于 NegEx 的否定药物不良事件过滤　对于否定药物不良事件的检测，采用基于语言规则的否定检测工具 NegEx。NegEx 是一个自然语言处理系统，曾用于出院小结中否定的医疗事件的检测。之前已有利用 NegEx 标注生物医学文本的研究[16]，并从出院记录中识别出医疗事件[17]。语义过滤过程的伪码描述如下：

伪码 4：

输入：具有一对相关药物和事件的关系实例 *i*

　　R（*drug*，*event*）＝*True*

输出：*T*（*drug*，*event*），*drug* 与 *event* 的关系类型

步骤：语义过滤算法（*drug*，*event*）

1. if *drug FAERS*. *drug*（ ）then
2. *indications*←*FAERS*. *indication*（*drug*）
3. if *event indications* then
4. return *T*（*drug*，*event*）＝药物适用症
5. for *rule* ∈ *NegEx* do
6. if *instance i matches rule* then
7. return *T*（*drug*，*event*）＝ 否定的药物不良事件
8. else
9. return *T*（*drug*，*event*）＝药物不良事件

3.5　报告源分类

为了减少患者报告的药物不良反应中的噪声和冗余，使用报告源分类过滤那些不依赖于患者实际体验的药物不良事件报告。网络健康社区中会出现一些来自第三方账号的与药物不良反应相关的新闻、故事、传说等复制或转载消息，这些消息不是患者实际的药物不良反应体验，之前的健康社交媒体研究没有关注过这个问题的解决。基于对已有研究的回顾发现，文本分类技术可以有效地识别雅虎问答中医学保健专业人员发布的医疗健康帖子，还能够从推文中识别出吸毒者，这些任务接近于从网络健康社区识别患者实际经历的药物不良事件[18]，基于统计学习的分类技术可以帮助从社交媒体中过滤掉这些噪音数据。

为了对社交媒体中药物不良事件的报告来源进行分类，采用基于特征的分类模型来区分患者的报告和传闻，利用词袋（Bag of Words，BOW）特征和直推式支持向量机（Transductive Support Vector Machine，TSVM）进行分类。采用词袋特征对数据集中的新闻、研究、传闻、复制等报告源进行特征选择，以区分患者实际经历的药物不良事件和传闻。TSVM 利用已标注的和未标注的数据构建模型，以一组小规模的已标注数据在未标注的数据中进行直推式推理[19]。

4　实验及结果分析

4.1　数据集及预处理

像糖尿病和心脏病等慢性疾病，常常依赖于患者的自我管理。许多在线健康论坛的出现，为慢性疾病患者提供了一个可以匿名交流的平台，在这里患者可以咨询问题、获取知识和分享自己面对疾病治疗中的喜怒哀乐。本研究中的实验数据集来源于美国著名的糖尿病社区 Diabetes Forums，论坛界面如图 4 所示。Diabetes Forums 是一个大型的糖尿病支持在线社区，拥有超过 50 000 的注册用户，网站上的大多数用户是糖尿病患者，还有一些是糖尿病护理人员。社区汇聚了关于糖尿病症状、治疗、监测、饮食和研究等的最新消息和讨论。



图 4　Diabetes Forums 界面

使用八爪鱼采集器爬取 Diabetes Forums 上从 2009 年 1 月 1 日至 2015 年 12 月 31 日的 67 444 篇贴文，由于本研究集中在句子层面信息的提取和处理，使用自然语言处理工具 OpenNLP 对每篇贴文进行断句后得到 42 355 个句子。

4.2　评价标准

采用标准的统计学习和文本分析评估指标：准确率、召回率和 F 值评估框架的性能，这些评价指标已被广泛应用于信息抽取和健康社交媒体的研究。

共现分析方法由于其简单易用常被用于抽取药物不良反应关系，因此其它改进的方法常常以该方法为基准进行对比。共现分析基于医学实体在文本中出现的概率来识别它们之间的关系。这种方法假设如果两个医学实体在一定范围内被同时提到，则它们之间存在潜在的生物医学关系。将本文提出的框架与共现分析方法进行比较，评估本文方法抽取 ADR 信号时的表现。

4.3　结果分析

4.3.1　医学实体识别　首先，利用 MetaMap 从健康社

交媒体数据集中识别 UMLS 医学概念,然后利用 FAERS 进行筛选,剔除那些未在 FEARS 中出现过的医学实体名,最后对前面保留下来的每一个医学实体,查询 CHV 得到其对应的用户偏好词汇,这些偏好词汇之前未被 MetaMap 识别,用这些用户偏好术语检索患者评论数据集,以扩展医学实体的抽取。表 3 展示了应用本文方法识别的医学实体结果。

表 3　医学实体识别效果

方法	实体类型	Precision (%)	Recall (%)	F1 (%)
本文的方法	药物	92.5	87.1	89.7
	医学事件	86.5	78.7	82.5

结果显示,利用本文方法进行药物实体提取的 F 检验值达到 90%,医学事件提取的 F 检验值达到 80% 以上。较好的性能表现主要归功于将网民保健用语、基于知识的过滤和 FAERS 药物安全数据库多数据源的结合。此外,由于糖尿病论坛讨论的药物和医学事件都是与糖尿病相关的,因而与那些具有不同背景、多样主题讨论的其它健康社区相比,通用术语的一致性更高,这个原因也会导致更高的性能结果。

通过对实验结果的分析可以发现,药物实体识别的错误主要产生于药名的拼写错误和简称;医学事件实体识别比药物实体识别表现出更低的性能,原因在于医学事件识别的主要错误来源于患者对医学事件更多的模糊描述。例如,患者描述"hypo-symptoms"和"a low"都指代 hypoglycemia(低血糖),而在实际抽取过程中却无法将这些模糊的描述识别出来。为了进一步提高性能,需要应用更先进的机器学习命名实体标注器。

4.3.2　药物不良事件关系抽取　为了进行药物-不良反应关系检测,从数据集中随机抽取 400 个句子,本文的方法专注于确定同一句子中的药物和医疗事件关系,而同一帖子中跨句子间的药物与医疗事件关系不在本研究之列。

基于现有知识库中的信息和临床专家的建议,根据前文方法对这些句子进行内容编码标注。句子中的每一对药物和医疗事件被看作是一个关系实例。由两名研究人员对这些关系实例进行标注,当两者出现不同意见时交由第三方裁决。标注的关系实例统计信息如表 4 所示:

表 4　标注数据集中的药物和事件关系统计

关系类型			不存在关系	总计
药物不良事件	药物适用症	否定的药物不良事件		
155	77	18	150	400

为了证明本文方法的有效性,与基于共现分析的药物不良事件提取方法进行比较。借鉴文献[11]的研究,如果一种药物与某种药物不良事件在一篇帖子中同时出现 20 次及以上,则被视为共现。

笔者比较了共现分析方法(CO)与基于统计学习的方法(SL)以及本文提出的融合统计学习与语义过滤的药物不良事件提取方法(SL + SF)。表 5 显示了三种不同方法提取药物不良事件的性能结果。

表 5　三种不同方法提取药物不良事件的性能结果

方法	Precision(%)	Recall (%)	F1(%)
CO	37.7	100.0	54.8
SL	64.2	60.4	62.2
SL + SF	78.6	60.4	62.2

比较结果显示,本文提出的方法可以显著提高药物不良事件提取的准确率和 F 度量。统计学习有助于提高准确率,同时导致召回率的下降;语义过滤进一步提高准确率,而对召回率不产生影响。本文方法的准确率比共现分析方法高出约 31%,F 度量值高出约 10%。共现分析方法的准确率主要取决于数据集的质量。由于用户在健康社区不仅讨论药物治疗的效果,还会叙述诊断、症状、药物的适应症、服药原因等内容多样的主题,在他们的讨论中有时可能涉及大量的药物名称,这导致共现分析方法的准确率偏低。然而,对于药物警戒研究,更准确地捕获 ADR 信号比获得大量虚假的报告更有意义。本文提出的方法可以提高从社交媒体抽取 ADR 信号的准确度,提高健康社会媒体药物不良事件报告的质量。

笔者还注意到,采用基于最短依存路径核函数的统计学习方法时,召回率会有所下降(从 100% 降至约 60%),这是由于长句的关系实例中检测关系的错误引起的。这些长句的关系表示在已标注数据中出现的次数较少,从而导致低的学习率和召回率。这个问题可以通过结合主动学习(一种机器学习形式)来解决,该方法可以决定哪些关系实例应该被标注以得到更好的抽取效果。

大量错误的药物不良事件不能通过共现分析的方法过滤掉,笔者提出的模型可以更有效地在社交媒体中抽取 ADR 信号,大大降低了社交媒体数据的嘈杂和冗余,提高获得患者报告药物不良事件的准确率。

5　总结与展望

社交媒体的出现为医疗保健数据的收集提供了新

的途径,应用自然语言处理技术从社交媒体中抽取患者报告的 ADR 信号对于改善药物警戒的临床和科学知识具有很大的潜力。然而,从社会媒体中提取患者报告的 ADR 信号仍然是医学信息学研究面临的重大挑战。

笔者开发了一个利用高级自然语言处理技术抽取患者报告的 ADR 信号的研究模型。该模型包括患者讨论的药物和事件的医学实体识别、药物不良事件关系抽取、报告源分类三个主要组成部分。药物和事件的医学实体识别采用基于多词典源的医学实体识别方法,应对社交媒体用户网络语言表达的多样化和口语化问题。采用基于最短依存路径核函数的统计学习方法抽取实体关系,然后采用基于医学知识与规则的语义过滤方法进一步提高药物不良事件关系抽取的准确度。最后,利用报告源分类区分患者实际体验的药物不良事件和传闻。为了评估所提模型的性能,通过收集糖尿病论坛上的数据对模型的有效性进行验证,结果显示该模型的每一部分都有助于其整体性能的提升。将模型应用于分析不同疾病的治疗和提取其它相关主题社交媒体的患者报告内容将是今后研究的方向。

参考文献:

[1] 梁少星,李枫林. 情景感知健康信息服务系统研究现状与展望[J]. 中华医学图书情报杂志,2014 (7):31 - 36.

[2] 王丹. 药品不良反应主动监测及其发展趋势[J]. 中国药物警戒,2015(10):600 - 602.

[3] HARPAZ R, DUMOUCHEL W, SHAH N, et al. Novel data-mining methodologies for adverse drug event discovery and analysis[J]. Clinical pharmacology and therapeutics, 2012, 91 (6): 1010 - 1021.

[4] LEAMAN R, WOJTULEWICZ L, SULLIVAN R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Berlin: Association for Computational Linguistics, 2010:117 - 125.

[5] BIAN J, TOPALOGLU U, YU F. Towards large-scale twitter mining for drug-related adverse event [C]//Proceedings of the 2012 international workshop on smart health and wellbeing. New York: ACM, 2012:25 - 32.

[6] CHEE B, BERLIN R, SCHATZ B. Predicting adverse drug events from personal health messages[C]//AMIA annual symposium proceedings. Bethesda: American Medical Informatics Association, 2011:217.

[7] 王丽伟. 药物不良事件信息资源整合与数据挖掘研究[D]. 长春:吉林大学,2014.

[8] ABEED S, RACHEL G, AZADEH N, et al. Utilizing social media data for pharmacovigilance: a review[J]. Journal of biomedical informatics, 2015, 54(C):202 - 212.

[9] THOMAS P, NEVES M, SOLT I, et al. Relation extraction for drug-druginteractions using ensemble learning[C]//Proceeding of the 1st challenge task on drug-drug interaction extraction. Huelva: Tranging, 2011:11 - 18.

[10] LIU X, CHEN H. A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports [J]. Journal of biomedical informatics, 2015, 58:268 - 279.

[11] BENTON A, UNGAR L, HILL S, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation[J]. Journal of biomedical informatics, 2011, 44(6): 989 - 996.

[12] YANG C, JIANG L, ZHANG M. Social media mining for drug safety signal detection[C]// Proceedings of the 2012 international workshop on smart health and wellbeing. New York: ACM, 2012: 33 - 40.

[13] 代菲,陈盛新,舒丽芯,等. 5 种信号挖掘方法在药物不良反应检测中的分析和应用[J]. 中国医院药学杂志,2012,32(20): 1674 - 1677.

[14] BUNESCU R, MOONEY R. A shortest path dependency kernel for relation extraction[C]//Proceedings of the conference on human language technology and empirical methods in natural language processing. Vancouver: Association for Computational Linguistics, 2005:724 - 731.

[15] LI J, ZHANG Z, LI X. Kernel-based learning for biomedical relation extraction[J]. Journal of the American Society for Information Science and Technology, 2008, 59(5):756 - 769.

[16] VINCZE V, SZARVAS G, FARKAS R, et al. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes[J]. BMC bioinformatics, 2008, 9(S11):1 - 9.

[17] UZUNER Ö, GOLDSTEIN I, LUO Y. Identifying patient smoking status from medical discharge records[J]. Journal of the American Medical Informatics Association, 2008, 15(1):14 - 24.

[18] LIU X, CHEN H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums [C]//Proceedings of the international conference on smart health (ICSH 2013). Berlin: Springer, 2013:134 - 150.

[19] JOACHIMST. Transductive inference for text classification using support vector machines[C]//Sixteenth international conference on machine learning. San Francisco: Morgan Kaufmann, 1999:200 - 209.

作者贡献说明:

魏巍:提出研究思路,组织和撰写论文;
郑杜:实施实验过程,进行数据收集及结果分析。

The Study of Adverse Drug Reaction Signal Extraction Framework Based on the Integrated Statistical Learning and Semantic Filter

Wei Wei[1] Zheng Du[2]

[1] Big data Institute, Zhongnan University of Economics and Law, Wuhan 430074

[2] The Center for the Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] The emergence of social media provides a new way to collect healthcare data. By using natural language management technology, the adverse drug reaction (ADR) signal can be extracted from social media, it has great potential to improve the clinical and scientific knowledge of ADR monitoring. However, the extraction of ADR from patients' reports in the social media is still a major challenge. This paper puts forwards an adverse drug reaction signal extraction framework based on advanced natural language processing techniques. [**Method/process**] The ADR signal extraction framework include the following implementation steps: Firstly, it recognizes the medical entity from the noisy social media based on multi-dictionary sources matching. Secondly, it applies statistical learning based on the shortest dependency path kernel to extract the adverse drug events. Then, filtering the information on the treatment and application of drugs as well as negative drug adverse events by though the semantic knowledge of the drug safety database. Finally, in order to remove rumors and other noise information, we should categorize the source of the report. [**Result/conclusion**] We collect data from BBS diabetes to identify the validity of the model, the result shows that each part of the model contributes to its overall performance.

Keywords: medical entity recognition adverse drug event extraction health social media statistical learning semantic filter